AFRL-IF-RS-TR-2006-138
Final Technical Report
April 2006

ALGORITHMIC TOOLS AND COMPUTATIONAL FRAMEWORKS FOR CELL INFORMATICS

New York University

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. M302/01

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-138 has been reviewed and is approved for publication

APPROVED: /s/

THOMAS RENZ
Project Engineer

FOR THE DIRECTOR: /s/

JAMES A. COLLINS
Deputy Chief, Advanced Computing Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
APRIL 2006

3. REPORT TYPE AND DATES COVERED
Final Sep 2001 – Sep 2005

4. TITLE AND SUBTITLE
ALGORITHMIC TOOLS AND COMPUTATIONAL FRAMEWORKS FOR CELL INFORMATICS

5. FUNDING NUMBERS
C - F30602-01-2-0556
PE - 61101E
PR - BICO
TA - M3
WU - 02

6. AUTHOR(S)
Bud Mishra

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
New York University
70 Washington Square South
New York New York 10012

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFTC
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-2006-138

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Thomas Renz/IFTC/(315) 330-3423 Thomas.Renz@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

The central thesis motivating this research project was that, by drawing upon mathematical approaches developed in the context of dynamical systems, kinetic analysis, computational theory and logic, it is possible to create powerful simulation, analysis and reasoning tools for working biologists. These tools could be used in deciphering existing data, devising new experiments and ultimately, understanding functional properties of genomes, proteomes, cells, organs and organisms. Through tool development (Simpathica, XSSYS, & GOALIE), analysis of important biological systems (apoptosis, purine metabolism, cell cycle, host-pathogen interaction, etc.) and foundational mathematical and algorithmic research (algebraic algorithm model checking, hidden Kripke model, best basis description, etc.), this project has succeeded in bring systems biology closer to its goals.

14. SUBJECT TERMS
Biocomputing, BioSpice, Bio Informatics, Proteomics, Cell Metabolism Analysis

15. NUMBER OF PAGES
30

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

Table of Contents

1. Introduction 1
2. Methods, Assumptions, and Procedures 2
   - Modeling and Analysis 2
     - Tools 3
     - Foundations 3
     - Models 5
   - Data Analysis and Redescription 6
   - Temporal Logic 8
3. Results and Discussion 9
   - Host-pathogen interaction SEB 9
   - Other Projects 10
     - Multi-frequency Analysis of Various Biological Systems 10
     - *C. elegans* Gonad Tract Cells Simulations 11
     - Multiple Synthetic Biological Models Classification 13
   - Tools and Systems 14
     - Simpathica/XSSYS 14
     - GOALIE 16
     - NYUMAD/NYUSIM 18
4. Conclusions 22
5. References 23

List of Figures

Figure 1: The Caspase Cascade pathway and its rendition with Simpathica...................... 5

Figure 2: A screenshot of GOALIE, a summarized result of the tests run with the Yeast Cell Cycle dataset and two TL formulae that can be derived from the information internally maintained by GOALIE. 7

Figure 3: The stem cell transition model, representing the possible changes that such a cell can undergo. 12

Figure 4: The 2D Java Cell Population Visualization Tool. 13

Figure 5 The Simpathica/XSSYS data flow schema. 16

Figure 6 NYUMAD/NYUSIM three tier architecture. 18

Figure 7 The NYUSIM client login window. 19

Figure 8 A screenshot of the NYUSIM interface. The main window shows a time course dataset obtained by running the Simpathica simulation front end. 20

Figure 9 NYUMAD/NYUSIM interface used to add experimental factors and other information to a given dataset. 21

1. Introduction

Biology thrives on complexity, and yet the approaches to decipher complex biological systems have been simple, observational, reductionist and qualitative. The observational nature of biology may even seem self-evident, as for instance expressed below more than three centuries ago by Robert Hooke, whose work *Micrographia* of 1665 contained his microscopical investigations that included the first identification of biological cells.

> "The truth is, the science of Nature has already been too long made only a work of the brain and the fancy. It is now high time that it should return to the plainness and soundness of observations on material and obvious things."

Nowadays, the fundamental "parts" out of which biology is created are observed, inferred and listed in multifarious ways. The results of these observations, inferences and cataloging are a marvel to regard as their interconnections, intertwining, and interactions appear to use some universal principles that still remain to be deciphered and fully understood. In order to unravel this biological complexity, it has become necessary to develop novel tools and approaches that augment, and rigorously formalize those human reasoning processes, which until now could be used for only tiny toy-like subsystems in biology. To this end, various "*Computational Systems Biology Tools*" aim to draw upon constructive mathematical approaches developed in the context of dynamical systems, kinetic analysis, computational theory and logic. The resulting toolkits aspire to build powerful simulation, analysis and reasoning facilities that can be used by working biologists for multiple purposes: in making sense of existing data, in devising new experiments and ultimately, in understanding functional properties of genomes, proteomes, cells, organs and organisms. If this ambitious program is to ultimately succeed, there are certain *critical components* that require special attention of computer scientists and applied mathematicians.

1) There is a critical need for powerful computational environments, where novice users can build prototyping tools quickly. An example of such a tool is the multi-scripting Valis environment, which provides rapid prototyping facilities in the same way as Matlab and Mathematica do for other disciplines. See [15].

2) There is a critical need for research and pedagogic modelling tools that allow a novice user to understand, reason, and ponder about large, complex and detailed biochemical systems effectively, efficiently and yet effortlessly.

3) Finally, there is a critical need for a catalogue of illustrating examples, where the afore-mentioned methodologies prove their power unambiguously.

Given the infancy of this emerging field, these pioneering experiments will face many unpredictable hurdles, but the experience gained will most likely revolutionize the current collective scientific viewpoint. Primary among these grand challenges could be the one related to various processes involved in cancer: cell cycle regulation, angiogenesis, DNA repair, apoptosis, cellular senescence, tissue space modeling enzymes, etc. Note that presently there is no clear way to determine if the current body of biological facts—in this instance, the ones related to cancer—is sufficient to explain phenomenology. In these particular cases, rigorous mathematical models with automated tools for reasoning, simulation, and computation can be of enormous help to uncover cognitive flaws, qualitative simplification or overly generalized assumptions.

2. Methods, Assumptions, and Procedures

The research program at NYU aimed to fulfill these critical needs described previously. The original development program described in the NYU proposal was quite ambitious, but did not foresee its relation to all the complementary work that began to be done within the BioCOMP program. Not surprisingly, the original plan went through a series of transformations as it proceeded and acquired new results produced by the laboratory, by other BioCOMP program members, and by the biology, bioinformatics, and systems biology communities at large.

The research carried on by the NYU Courant Bioinformatics Group developed along two complementary lines sharing many ideas, approaches and concepts, with several sub-projects supporting them.

- **Recomputation**: The first research line concentrated on *modeling* and *analysis* of biological systems via simulation and "formal methods".
- **Redescription**: The second research line concentrated on the direct analysis of measured data (mostly micro-array data). Statistical approaches and representation problems constituted the core of this research line.

Each research line carried its own set of peculiar methods, assumptions and procedures. However, a unifying effort was always pursued during the project development.

Modeling and Analysis

One of the key objectives of the NYU Courant Bioinformatics Group has been to provide biologists with new tools capable of aiding the formulation of alternative hypotheses, and thus, reducing the effective cost of an experiment. Model building and model checking tools were considered the most ideal for this goal.

The modeling of biological systems relies on several techniques developed over the years. Deterministic and stochastic methods (e.g., the well known Gillespie's algorithm) have been used for a variety of examples in literature, in conjunction with different modeling formalisms and tools. In particular, *hybrid systems* [1] have proved well suited for the modeling needs of systems biology, as they allow for a combination of discrete and continuous semantics, which can be used to model and approximate both regulatory and metabolic networks. Furthermore, the study of hybrid systems required (and still requires) answering deep and complex mathematical and algorithmic questions, to address issues of scale and efficiency that hamper the direct observation of large biological system models.

Tools

In order to complete the research program, the NYU Bioinformatics Group built several tools that have been distributed under the DARPA Open Source License and that have been – when feasible – integrated into the BioSPICE framework.

Simpathica is a toolset that addressed the issues at the core of the NYU Bioinformatics Group statement of work. Simpathica comprises an Ordinary Differential Equations (ODE) based pathway simulator, which is used as the main engine for producing different scenarios by varying the model's parameters. The simulation engine is based on off-the-shelf ODE integrator libraries (Octave LSODE library and Python's NumPy and SciPy libraries), enhanced with features that are proper for a hybrid system simulator.

Simpathica also comprises a "simulation traces analysis" tool based on a query language patterned after a well-known temporal logic. Given a set of simulation traces, a user can formulate a number of "temporal queries" about the behavior of the system. The Simpathica temporal logic analysis back-end (called XSSYS) sorts through which query is true and which is not. This kind of analysis is complementary to the usual charting of simulation results, and it has two advantages. First, it can be easily automated, as queries and data are all computer readable. Second, it can find small discrepancies in the data, which may escape simple analysis of graphed values, especially in the presence of several variables.

GOALIE is a software system that uses the GO ontology biological process taxonomy to explore temporal invariants, directly interpreting numerical data organized along time (or concentration, dosage or other independent variable, or combination thereof). The key contribution afforded by GOALIE is integrating data-driven reasoning about time course datasets (mostly micro-array data) with model-building capabilities through the concept of *redescription*.

NYUMAD and NYUSIM are two databases that were built to support the data storage needs of the NYU Bioinformatics Group and its collaborators. NYUMAD is a database for storing micro-array data based on the MAML model definition. NYUSIM is a database system for storing simulation data.

Foundations

Several inquiries and improvements were also made on the theoretical and foundational aspects of hybrid systems analysis. In conjunction with other researchers in the BioCOMP program, the NYU Bioinformatics Group expanded the algorithmic and mathematical foundation of hybrid systems analysis tools, by re-casting the key problem, namely *reachability analysis*, in terms of *bounded semi-algebraic model checking*.

Experiments and simulations produce *traces* of observable quantities of various biochemical systems at different levels of detail and accuracy (i.e., *time-course* data.) Acquiring and integrating several data sources, e.g., from *in vivo*, *in vitro*, and *in silico* (i.e., simulated) experiments, and generating one or more observed traces is thus an important activity. The data collected is stashed in various databases and formats. E.g., the NYUMAD database stores micro-array data and allows for collaborative efforts across geographically distributed laboratories.

The data collected along with the models of the biological system being studied—e.g., the Caspase cascade system—constitutes the primary elements of our analysis. The models may be selected from a variety of sources in an interoperable manner, e.g., SBML, BIOCYC, WIT, KEGG, PathDB, etc.

Once we have the data and a model, our system constructs a representation of the "regions of interest" of the parameter space which satisfy a given *reachability* condition. As an example, we could identify for which ranges of parameters the system reaches a steady state. On account of the symbolic algebraic manipulation of the models, we can describe concisely and exhaustively such "regions of interest." Once these regions are available for analysis, it will be possible to plan the wet-lab experiments around them. While it can be argued that this is not all that different from current practice, our symbolic methods provide exhaustiveness, scalability, and correctness when combining different models to explore emergent behavior.

A survey of several pathway models in Systems Biology reveals that a vast majority of the equations presented can be recast in a relatively small set of *canonical forms*. In particular, several of basic mathematical models for metabolic, regulatory and signaling pathways—Michaelis-Menten, Hill type *cooperative* equations, *Generalized Mass-Action* models—can be accommodated by casting them into an appropriate hybrid system formalism with a few constraints.

Essentially, the behavior of the system can be modeled as a set of equations which in this case *could* be the model of a metabolic pathway expressed as a set of *mass-action* equations of the (restricted) form

$$\dot{x} = \frac{P(x_1, x_2, ..., x_k; t; p_1, p_2, ..., p_k)}{Q(x_1, x_2, ..., x_k; t; p_1, p_2, ..., p_k)}$$

with the 'x's denoting the system observable variables, t denoting time, and the 'p's being the *parameters* of the model, i.e., a *rational* function of *polynomials* P and Q. Tarski's theorem and related decision procedures (e.g. Collins' cylindrical algebraic decomposition (CAD) for quantifier elimination) are therefore applicable, and can be used to compute the reachability region of the model representing the biological system.

This encoding has advantages from the point of view of purely *symbolic* manipulation of the kind carried out while doing paper-and-pencil algebraic manipulation. The reasons are twofold. First, the manipulation of rational functions is the most developed one in the field of *Computer Algebra*. Second, while it is true that the algorithms for the manipulation of rational functions are very time consuming in the worst case, we note that such complexity is dependent on the "exponents" appearing in the polynomials. For the case of biochemical processes we are interested in modeling, these exponents are low: usually, below 3, and much more rarely 4. This reduced complexity in biological examples makes the symbolic manipulation algorithms applicable to much larger systems than in the unrestricted case.

The primary result of this line of inquiry has been the precise characterization of the subclass of hybrid system models apt for biological system modeling. Further, appropriate approximation schemes have been proposed, that can be manipulated by a program in a semi-automated way, while bounding the inherent computational complexity of the approach. (See [2-15] for a detailed exposition)

Models

Several biological systems were selected for analysis in consultation with biologist collaborators, in order to test the viability of the approaches and the techniques developed under the BioCOMP program.

Caspase cascade models. The NYU Bioinformatics Group and the Lazebnick's laboratory of Cold Spring Harbor Lab have completed several models, consistent with predominant hypotheses about how internal-signal based apoptotic processes operate to cause cell death. The models of interest are initiated by cytochromes from mitochondria in response to DNA damage, and are also involved in creating a holoenzyme to degrade DNA and proteins in the cell. The holoenzyme creation process is aided by a Caspase-cascade, and the details of the process were firmly established by means of the models constructed and by the experiments performed, (Figure 1). The results of this analysis and its extensions to a similar process initiated by external-signaling have also been the subject of a related study on host-pathogen interaction that is described elsewhere in this report.

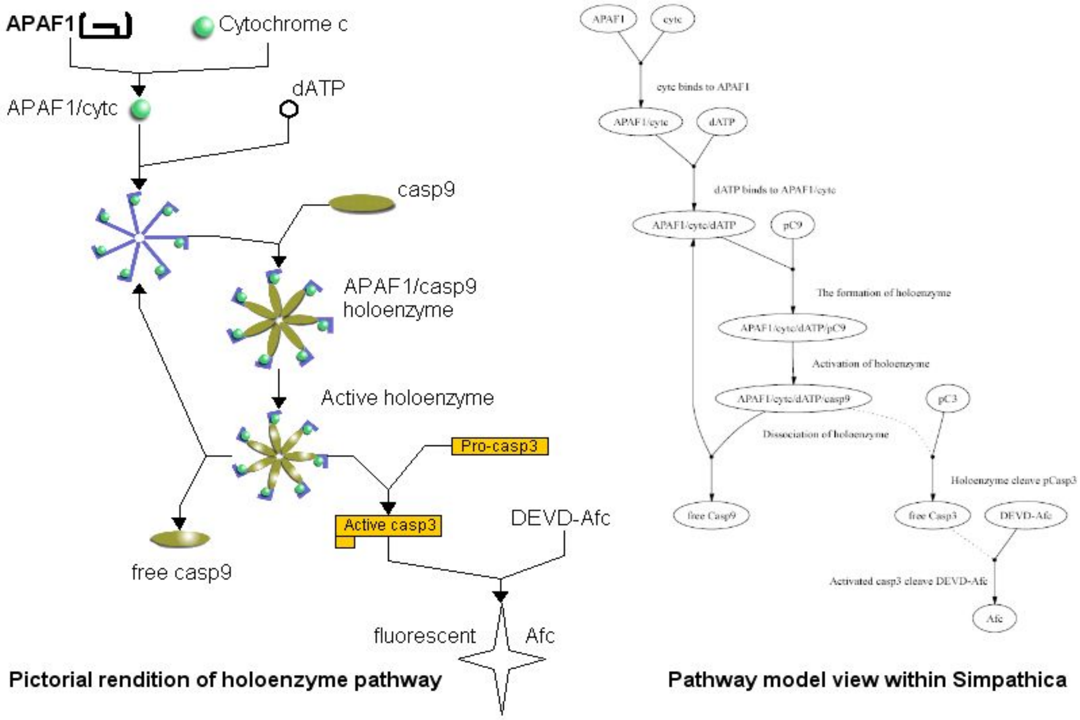


Figure 1: The Caspase Cascade pathway and its rendition with Simpathica.

SEB host-pathogen interaction. The SEB host-pathogen interaction case study was analyzed in collaboration with several groups within the BioCOMP program: Walter Reed Army Institute of Research, Thomas Jefferson University., and University of

California at Los Angeles. The pathogen attacks with *Staphylococcal enterotoxin B* (SEB), a member of a family of exotoxins produced by Staphylococcus aureus. It is a causative agent of toxic shock characterized by acute vasodilatation leading to severe hypotension. The SEB host-pathogen interaction case study was the test bed and the main motivation used to develop GOALIE.

Data Analysis and Redescription

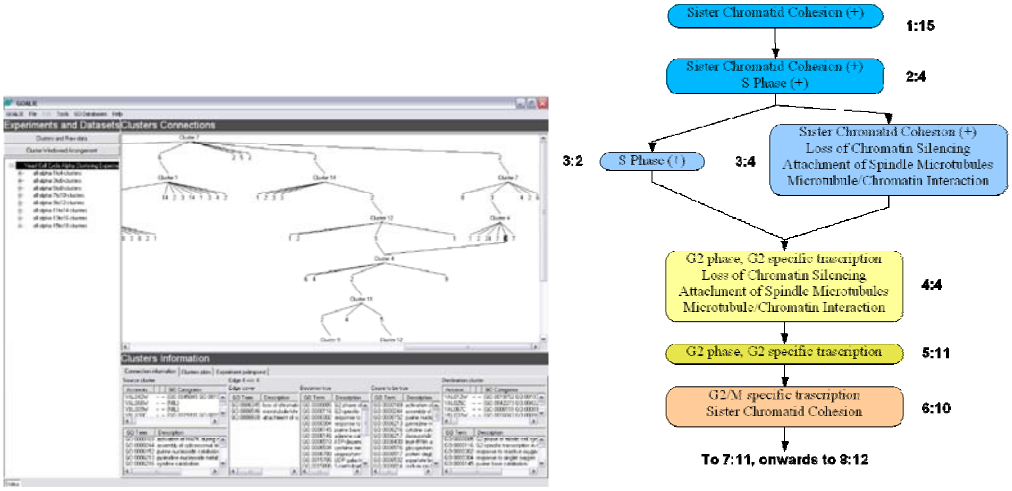
Biological processes are the result of interactions within complex networks. Traditionally, biologists constructed "models" to demonstrate their ideas on how a particular biological system or subsystem actually works, by relying on their immense breadth of knowledge, compounded by depth and expertise on a particular system. Now, biologists are also faced with the task of *reconstructing models* from large data sets, where relevant information may be deeply buried in layers of numerical information.

Current micro-array data analysis techniques draw the biologist's attention to targeted sets of genes e.g., those that vary in a well correlated manner, are under similar regulatory control, or that have consistent functional annotation or ontological categorizations. Yet, such methods do not present global or dynamic perspectives (e.g., invariants) inferred collectively over the dataset. Such perspectives are important in order to obtain a process-level understanding of the underlying cellular machinery, especially how cells react, respond, and recover from stresses.

To address these problems, an approach and a system based on the statistical analysis of time course micro-array data were developed, by taking into account the annotations that have been made of each gene, which are available from the several publicly available databases. The procedural ideas at the core of such an approach are relatively simple, yet they have already opened up a number of very interesting mathematical questions regarding the overall performance of the setup, and the way its results can be interpreted by a biologist, especially in the face of very noisy data and irrelevant pieces of information. The end result is to further annotate an experiment with "invariant properties" constructed from the building blocks provided by a controlled vocabulary. Expressing such invariants using Temporal Logic (i.e. one of its variants) appears to be the most natural step to take.

To recapitulate, on one side there is a growing number of (time-course) datasets from micro-array experiments, while on the other side, there is a growing corpus of gene products and protein annotations in terms of controlled vocabularies (or ontologies). Several researchers have at this point started to look at the significance of "describing" a micro-array dataset with the terms contained in such ontologies, e.g. the Gene Ontology, the MeSH or the BioPAX ontology. In these applications, the question asked is often of the form "what is the set of terms that conveys the most information about an experiment?". This question is usually asked after a "clustering" of the data is performed and appropriate statistical inference tests have been applied (e.g. a Fisher Exact Test).

The approach developed within the BioCOMP program by the NYU Courant Bioinformatics Group adds one more dimension to the analysis, by introducing a breakdown of the time course experiment (which usually contains from 5 to 50 time points) into “overlapping windows”. Clustering is performed within each window and the biological processes are tracked as they “move” from window to window. This yields valuable information to a biologist about the locality of phenomena, which may be the subject of better-targeted – hence cheaper – experiments. The result is a set of graph relationships between windows based on the associations among clusters and terms from the controlled vocabulary. This set of graph-relationships is the basis for the construction of temporal logic formulae describing the biological system at a phenomenological level. The construction of this graph is straightforward but it strongly depends on the choice of controlled vocabulary or ontology, on the quality of the basic annotations available (e.g. annotation of a given gene product with a number of terms), and on the quality of the statistical tests used to perform the initial association of ontology terms to the clustering experiment in each window, (Figure 2).


Experiments and Datasets
Clusters Connections
Clusters Information
Sister Chromatid Cohesion (+)
1:15
Sister Chromatid Cohesion (+)
S Phase (+)
2:4
3:2
S Phase (!)
3:4
Sister Chromatid Cohesion (+)
Loss of Chromatin Silencing
Attachment of Spindle Microtubules
Microtubule/Chromatin Interaction
G2 phase, G2 specific trascription
Loss of Chromatin Silencing
Attachment of Spindle Microtubules
Microtubule/Chromatin Interaction
4:4
G2 phase, G2 specific trascription
5:11
G2/M specific trascription
Sister Chromatid Cohesion
6:10
To 7:11, onwards to 8:12

```
Exists_path(`sister chromatid cohesion'
            Until (`G2 phase' And `G2 specific transcription'))
Eventually(Exists_path((`G2 phase' And `G2 specific transcription')
                       Until `G2/M specific transcription'))
```

Figure 2: A screenshot of GOALIE, a summarized result of the tests run with the Yeast Cell Cycle dataset and two TL formulae that can be derived from the information internally maintained by GOALIE.
The GOALIE display is divided in two parts. The left part contains a list of all the clusters that are part of the analysis. The right part, split into top and bottom views, contains information about the relationships among the different clusters. The top view shows the “graph” of connections among clusters at different time points; two clusters being “connected” if they share some characteristics i.e. GO categorizations. The bottom view shows information about the connection between two clusters: which GO categories are maintained, which are present in the first but not in the second, and which are present in the second, but not in the first.
GOALIE has all the pre-processed information available to automatically generate these two temporal logic formulae. The first one states that there exists a directed path connecting a sequence of clusters in successive time windows such that the GO category

`sister chromatid cohesion' holds until *the cell enters G2 phase. The second formula states, albeit obviously, the following: "the cell, after dwelling in G2 phase, enters M phase" Although this is a well known feature of the cell cycle, it is interesting as it derives automatically from numerical expression matrices and a static ontological annotation*

Finally, the set of graph relationships is organized in a directed acyclic graph (DAG), although circularities can be re-introduced by a wrapping technique in order to analyze periodic processes, e.g., in studies of the Cell Cycle. An edge is placed between a cluster in a window and another cluster in a previous or successor window. Each edge is tagged with the terms that are shared between the two clusters' re-descriptions. Each edge is also tagged with the terms that are associated only to the first cluster, and with the terms that are associated only to the second cluster. The set of temporal logic sentences is reconstructed by analyzing different "chains" of edges in the DAG. E.g. finding a set of terms that appear in each edge of a chain from the initial window to the last window generates a particular temporal logic sentence, denoting the invariance of the set of terms.

The set of invariants, whose semantics is well defined in terms of an interpretation of the DAG as a Kripke Structure, can be rendered using appropriate linguistic interfaces. A biologist reading such a description will be able to discern particular phenomena, which may emerge from the analysis of the data, and design future experiments accordingly, thus lowering their cost.

Of course, different ontologies will produce different descriptions of the dataset. Hence the question: when are two descriptions comparable? Assuming that the results of a biological experiment cannot have wildly diverging interpretations, two different controlled vocabularies should produce "compatible" descriptions. Deciding that the two descriptions are compatible will require the application of several linguistic techniques. However, should the descriptions diverge in a significant way; this difference could point to an unresolved scientific problem. For instance, this divergence could point a biologist toward the design of a targeted experiment whose aim would be to resolve the conflict. The design of the experiment will be dependent on the quality of the description made in terms of the different ontologies; thus, the need for appropriate linguistic interfaces.

Temporal Logic

A common trait of all the research carried on was the use of one form or another of *Temporal Logic (TL)* as a canonical representation of properties of biological systems. The temporal logic chosen was at a time a variation of computation tree logic (CTL) or a variation of a *linear* temporal logic (LTL).

Temporal logic was used to formulate queries to be tested against simulation traces in Simpathica. Simpathica also contains a Natural Language interface that transforms English questions into appropriate TL queries for the trace analysis system.

Temporal logic formulae can also be generated in a controlled way to produce "descriptions" of a set of simulation traces or of a time series of micro-array experiments. GOALIE can produce such a set of TL formulae by leveraging the redescribed clusters and their inter-relationships. Once GOALIE has produced the TL formulation, rendering it in a simple set of English sentences is a straightforward task.

3. Results and Discussion

The research activities of the NYU Bioinformatics Group led to several results and tools. This section is roughly organized into two parts: the first one describing the biological systems that were analyzed, while the second part briefly describes the tools that were developed during the funding period.

Host-pathogen interaction SEB

In collaboration with several other groups within the context of the DARPA BioCOMP program, and with Dr. Jett's group at Walter Reed Army Institute of Research in particular, the NYU Bioinformatics Group analyzed a host-pathogen interaction dataset. The pathogen attacks with *Staphylococcal enterotoxin B* (SEB), a member of a family of exotoxins produced by *Staphylococcus aureus*. It is a causative agent of toxic shock characterized by acute vasodilatation leading to severe hypotension. Animal studies have shown that 75% of the toxin administered to primates localizes to renal proximal tubule epithelial cells (RPTEC), which possess a glycosphingolipid receptor for SEB and express apoptotic markers in response to the toxin. These cells also secrete potent vasoconstrictors, endothelins, which play an important role in vascular tone regulation [16-19]. A highly collaborative research effort within the DARPA program was aimed to reconstruct a model of the pre-apoptosis processes involved in the response to SEB.

The GOALIE tool was used to analyze the SEB dataset provided by WRAIR Jett's Laboratory. The dataset comprises several time course micro-array measurements of gene expression levels under two treatments. The observations made on a dataset prepared from the 50μg treatment set of more than 700 genes are described below.

The time course data was partitioned in windows of 3 time points, yielding 4 windows. Each window was partitioned into 20 clusters giving a total number of 80 clusters. These 80 clusters were redescribed at a *p*-value of 0.05; the redescriptions across windows were then computed using a stringent Jaccard's coefficient $\theta = 0.8$.

The notation: `*L:N*,' with *L* and *N* positive integers, to denote cluster *N* in time course window *L* is used in the following paragraphs

Time Course Window 1 to Time Course Window 2: Connection 1:9 to 2:18. By inspecting the first cluster in the first window (Cluster 1:9), it can be noted that one of the connections to a cluster in the second window (Cluster 2:18) is labeled (among many others) by the GO categories “circulation” (GO:0008015), and by the category “negative regulation of heart rate” (GO:0045822). This represents a constant set of biological processes shared by this cluster chain, traversing Cluster 3:17 to Cluster 4:13.

Time Course Window 1 to Time Course Window 2: Connection 1:9 to 2:6. The connection between Cluster 1:9 and Cluster 2:6 is interesting because it shows how the category “regulation of lymphocyte proliferation” (GO:0050670) becomes activated in the next time-window (Cluster 2:6), while the categories “antigen presentation” and "antigen processing" become inactive. This should indicate that some of the genes in the clusters start a response to the pathogen in the second time point.

Time Course Window 2 to Time Course Window 3: Connection 2:6 to 3:3. Following the chain downward it can be noted that the lymphocyte proliferation activity is maintained, while categories "lymphocyte differentiation" and "B-cell differentiation" become active in Cluster 3:3. This suggests another stage in the response of the cells to the pathogen.

Following the chains downward, other interesting categories that have been maintained across the time course are evident. E.g., there is a set of genes involved in the "cell adhesion" processes (chain starting at Cluster 1:16, following 2:8, 3:20 and 4:19) which may suggest some "mechanical effect" being at play.

Another interesting set of relationships involves Clusters 3:3 and 3:6 in two different chains rooted at 1:9 and 1:2. The two clusters connect to cluster 4:3 because of two separate, yet apparently related groups of categories. The connection 3:3-4:3 is labeled by the inactivation of the "B-cell differentiation", "lymphocyte differentiation", and "regulation of lymphocyte proliferation" categories. The connection 3:6-4:3 shows instead the inactivation of the "cellular defense response (Sensu Vertebrata)". These inactivations may be related at a deeper level as they are tied completely to sets of Open Reading Frames (ORFs) in the two clusters.

Other Projects

Multi-frequency Analysis of Various Biological Systems

The algorithm proposed was designed to analyze data obtained from time-course experiments (e.g., transcriptomic or proteomic data) or *in silico* simulation of biological processes. The input to the algorithm consists of trajectories for the dynamic evolution of the abundance of various molecules in a biological system, generated at different experimental conditions. The goal of the analysis was to determine whether variations in the experimental conditions (e.g., initial conditions or duration of stimuli) cause the system to evolve globally in a substantially different manner. Different modes of operation in the system could be identified and a correspondence between the typical experimental conditions and these modes of dynamic behavior could be established. For example, this technique was able to detect the differences in the evolution toward the two stable states of the Ras–Protein Kinase C–mitogen-activated protein kinase (MAPK) bistable pathway activated by EGF stimuli of various strengths. However, the differences in dynamic behavior that could be detected were not at all confined to multistable systems.

The simple mathematical observation underlying our approach was that it would be possible to choose a small number of vectors in an orthonormal basis, so that all the trajectories of the system under consideration are effectively described mostly by the coefficients with respect to those vectors. In mathematical terms, the characteristics of the set of trajectories of a complex biological system were studied by projecting them onto a suitable, low-dimensional vector space. Because any trajectory can be projected onto this ‘‘coefficient space’’ (more formally, the D-Space), it is then possible to project a large number of randomly sampled trajectories into points in the D-Space and identify the different modes of evolution of the system by inspecting the clusters that these projected points form in D-Space. Studying the geometric properties of these projected points can lead to the identification of the biological system modes. A more formal explanation of such techniques is given in [20].

C. elegans Gonad Tract Cells Simulations

Understanding intercellular signaling mechanism in developmental biology is an important task. In this context, several experiments on the nematode *C. elegans* were conducted in cooperation with colleagues in the NYU Department of Biology, in order to test several hypotheses regarding germ line development. The goal was to better understand the processes involved in stem cell differentiation and proliferation. No animal research was conducted under this project.

To this end, a rigorous computational model of *C. elegans* germ line stem cell growth, based on real observations of cell division patterns in the distal mitotic zone, provided information about the underlying signaling processes. Since the cell division patterns are not directly obtainable from live animals, synchronized populations of worms were fixed and examined for the presence of actively dividing cells within the distal mitotic zone. A large collection of such data was treated as a set with certain probabilistic parameters regarding the rate and position of cell division.

As part of the investigation of different approaches to simulation, a stem cell population model was developed, which was tested on a number of available simulation platforms. The aim of this study was to evaluate each simulation platform with respect to its ease of use and expressivity when dealing with the chosen problem.

The Stem Cell model is an augmented *Birth-Death* process. A population of Adult Stem Cells is seen differentiating into a set of *Committed Progenitors* Cells, which then differentiate into *Specialized Somatic* Cells. This model captures the behavior of stem cells. A picture of the complex transitions possible for a single Stem Cell follows, (Figure 3).

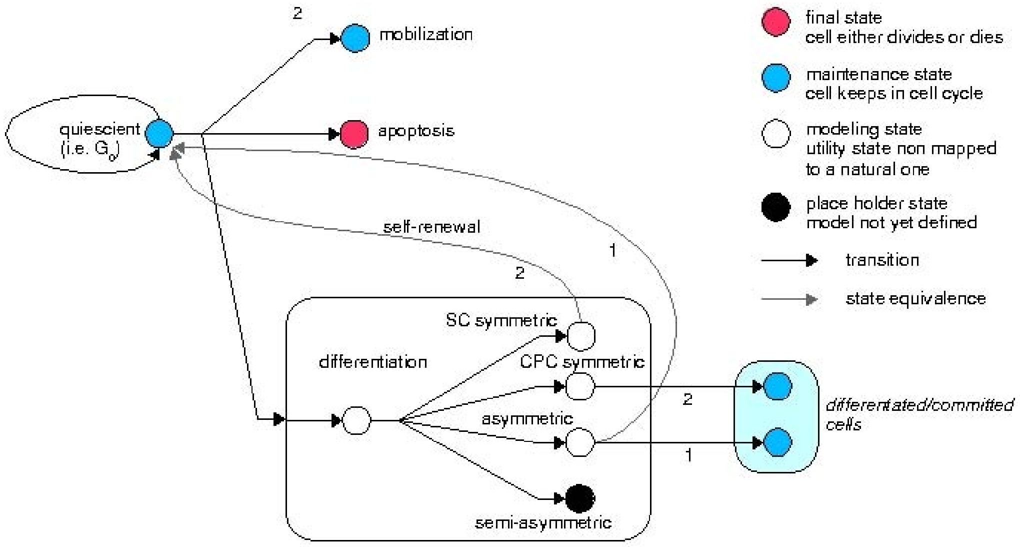


Figure 3: The stem cell transition model, representing the possible changes that such a cell can undergo.

The model consists of an approximated set of deterministic Ordinary Differential Equations, ODEs, modeling the stochastic transitions of the system. The model has several parameters, which are essentially the *rates* of the exponential processes representing each possible transition for a class of cells (apoptosis, differentiation). The sizes of the populations of different kinds of stem and somatic cells were logged as a result of a simulation run. The model was encoded as (1) a plain set of ODEs and as (2) a Discretized Stochastic Finite Automaton, which was then implemented in a number of environments: Octave (an ``open source" non commercial system very similar to MatLab), Mathematica, LambdaSHIFT (a Hybrid System simulator from UC Berkeley), and Charon (another Hybrid System Simulator from U. Penn), and (3) as a *spatially distributed* population of locally interacting cells.

The modeling of the Stem Cell Population system eventually evolved into a modular approach, where each stage of the differentiation process was distinguished simply by different parameters. The resulting model is rather robust, converges to the expected steady state, and responds well to perturbations in all the implementations.

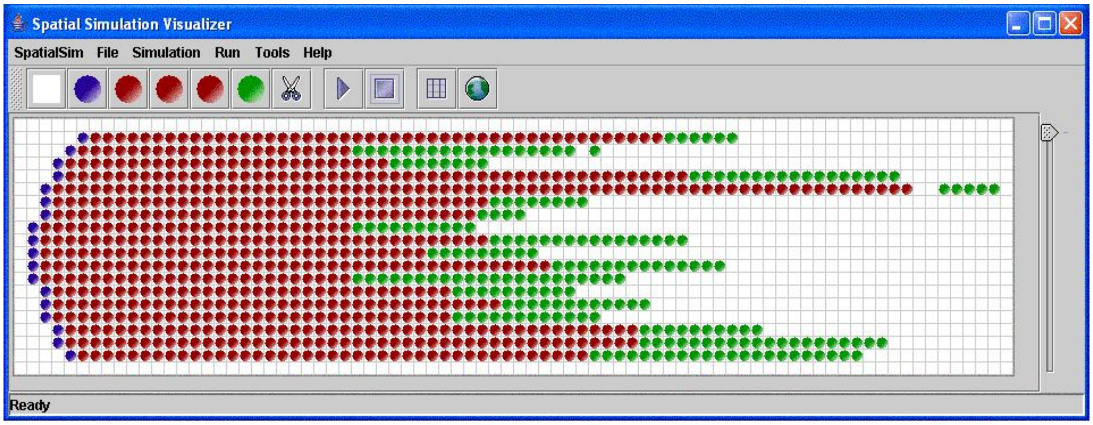


Figure 4: The 2D Java Cell Population Visualization Tool.

The initial model was an *aggregate* one, in the sense that it considered populations of cells simply by tracking their sizes. The *spatially distributed* model reuses the original set of equations to define the local rules of interactions among the different cells. A simple 2D/3D visualization tool that shows the evolution of a population of stem cells was developed to facilitate the debugging of the model, (Figure 4).

Multiple Synthetic Biological Models Classification

The systematic application of time-frequency analysis and temporal logic based model checking led to the formulation of a procedure for the classification of families of biochemical pathways and circuits in terms of their temporal behavior. There were two immediate pay-offs of this approach. First, a family of models obtained by systematic perturbations of an archetypal (e.g., wild-type or ancestral) model, when classified in this manner, can help in identifying various incorrect or implausible features of the model. Second, specific hypotheses about various features of a given model can be automatically generated from such analyses, and then subjected to more exacting experimental verification.

The approach was used (1) to understand the behavior of any individual topologically distinct circuit among the set of 125 synthetic biological circuits created out of similar elements, and (2) to ascertain the correctness of a well known Yeast Cell Cycle model by checking a family of perturbed models created through single and double mutations. The resulting tool interacts with a propositional temporal logic model-checking system to present qualitative distinctions among the groups within the family of biological circuits or among the different multi-modal behaviors of a single pathway.

The method was tested on the interesting problem posed by the analysis of the series of experiments of Leibler and Elowitz [21] and Guet et al. [22], in which the authors design and implement in-vivo a family of combinatorial circuits. The long term goal is to solve the problem of mapping and reconstructing a mathematically sound and complete model to a set of wet-lab observations. In the specific case of Guet et al.'s 125 combinatorial *in vivo* circuits, we want to be able to map the behavior of each of them to a model representing one of the standard Boolean gates.

The original motivation for designing such a family of synthetic networks by combinatorial variations of the network topology was given as follows [22]: "*A central problem in biology is determining how genes interact as parts of functional networks. Creation and analysis of synthetic networks, composed of well-characterized genetic elements, provide a framework for theoretical modeling. ... Combinatorial synthesis provides an alternative approach for studying biological networks, as well as an efficient method for producing diverse phenotypes in-vivo.*" Nonetheless, lack of efficient tools for modeling and analysis of such synthetic networks has hindered many possible applications of these networks. With appropriate tools, one can foresee applications where millions of randomly generated networks could be screened for selection of primitive circuits with specific properties (robustness, immunity to noise, etc.), or as building blocks of larger circuits with specific temporal properties, or even as scaffold structures for measuring kinetic parameters of a component as it operates in vivo. The *Simpathica/XSSYS* system and the ancillary modules *NYUSIM* and *NYU BioWave* respond to these demands quite well. That is, the combination of Simpathica modules can be employed synergistically to analyze the set of Guet et al.'s biological combinatorial circuits, by providing a semi-automated way to classify them based on the profiles of their behaviors. The classification of behaviors was accomplished by a careful application of time-frequency clustering techniques and by a modified model-checking approach that directly tests for the truth value of Boolean expressions over traces of the system. Following such demonstration, an extension of this approach was developed to search for important distinguishing logical characterization, by using a *generate-and-test* procedure for temporal logic formulæ. Such a process, though computationally inefficient, automates a *discovery system* that can further aid a working biologist. The approach was also tested against a model of the Yeast Cell Cycle [13,14], for which a preliminary group of three datasets had been produced [23].

Tools and Systems

Simpathica/XSSYS

The *Simpathica/XSSYS* system is a novel *Pathway Simulation and Trace Query* tool. The system eventually manipulates *traces*, which can be the product of wet-lab experiments or computer simulations.

Wet-lab experiments are extremely costly. Running simulated experiments before setting up a complex wet-lab experiment may guide the researcher toward a solution (verifying a hypothesis) in a quicker and more cost-effective way.

No matter what the source of data is, for most biological systems the number of variables involved is very high. It then becomes very difficult to concisely formulate queries about the system behavior.

The Simpathica/XSSYS system allows a user to construct a model of a set of pathways in the spirit of what was proposed by Voit et al. The set of pathways is translated into an XML format, which encodes a set of relations among products and enzymes and gives rise to a set of *Differential Algebraic Equations* (DAE, *S*-System, *cfr.* Voit [24]) with special constraints.

The set of DAE is simulated using Octave and the result is stored in an intermediate format suitable for the trace analysis subsystem. The resulting trace is eventually stored in a Database system, (Figure 5).

Given the trace of a system, i.e. a *time-indexed* sequence of state vectors representing a solution of the DAE system, Simpathica can perform the following operations.

Simpathica extracts a *collapsed* timed automaton from the trace by grouping the state vectors according to several criteria. The most important criterion is the detection of *significant changes* in the vector field underlying the DAE numerical solution. An equivalence relation based on a bisimulation analysis over the collapsed states is also built, in order to be able to detect cycles in the collapsed automaton (essentially achievable with a normalization operation).

The definition of a query language (Simulation Runs Query Language, SQRL) based on a Temporal Logic formalism was developed in which to formulate queries like

```
eventually(not always(LacI < 1.3) or always(LacI > 4.0)).
```

The above example query expresses the fact that the value of the `LacI' variable oscillates between the two values of 1.3 and 4.0. The system being analyzed is the `repressilator' system by Elowitz and Leibler [21]. The analysis tool provides counter examples of when a query such as this is not true and describes under which altered conditions it could be true.

The approach is based on the observation that the traces being analyzed do not necessarily represent all possible behaviors of the metabolic and regulatory system (i.e. of the set of DAEs) under scrutiny. This is a key observation that allows us to tailor the algorithms for much more efficient execution than in the case of general analysis and verification tools.

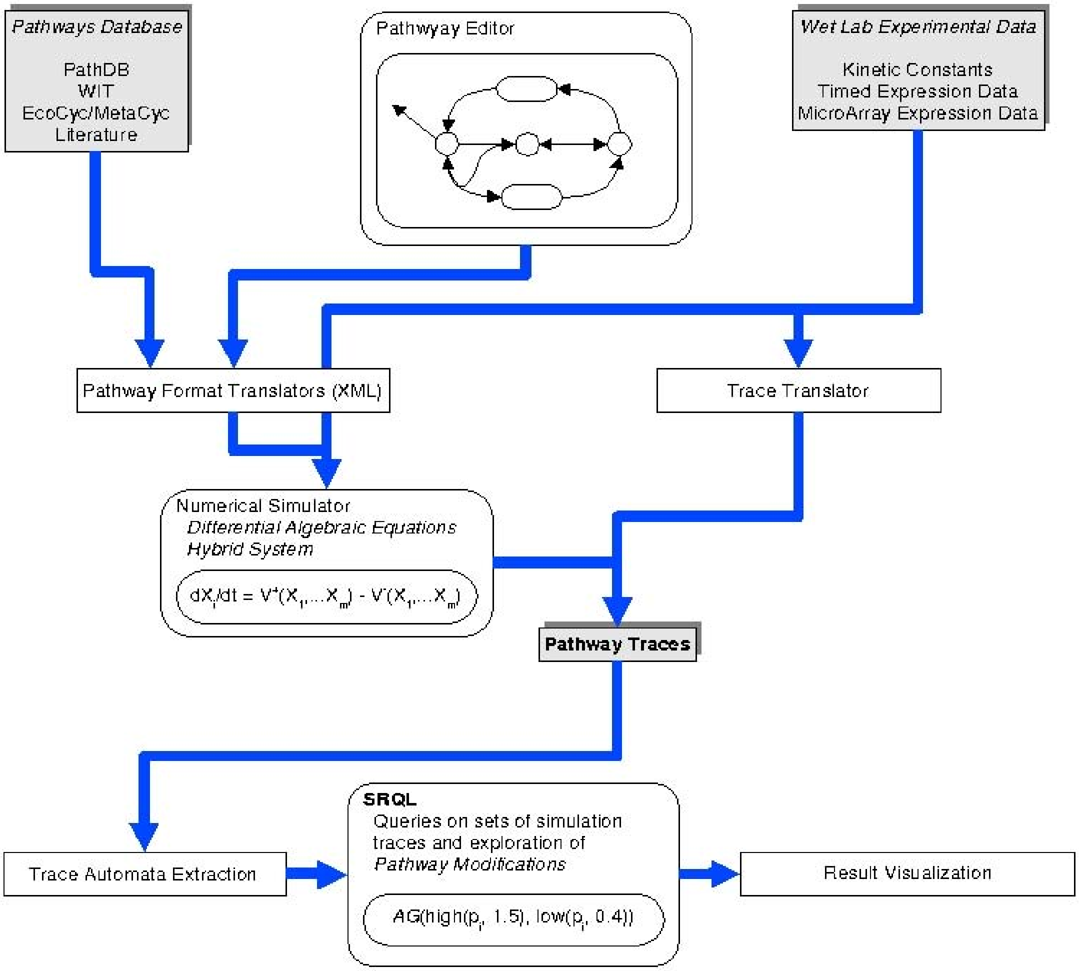

Pathways Database
PathDB
WIT
EcoCyc/MetaCyc
Literature
Pathwyay Editor
Wet Lab Experimental Data
Kinetic Constants
Timed Expression Data
MicroArray Expression Data
Pathway Format Translators (XML)
Trace Translator
Numerical Simulator
Differential Algebraic Equations
Hybrid System
dX_i/dt = V+(X_1,...X_m) - V-(X_1,...X_m)
Pathway Traces
SRQL
Queries on sets of simulation traces and exploration of
Pathway Modifications
AG(high(p_i, 1.5), low(p_i, 0.4))
Trace Automata Extraction
Result Visualization

Figure 5 The Simpathica/XSSYS data flow schema.

Moreover, while the analysis of a single trace roughly corresponds to the analysis of a *Linear Time Temporal Logic* system, *Branching Time TL* can be used to cross analyze several traces of a system. Thus Simpathica achieved a nice balance between expressiveness and efficiency for the task of comparing several traces in order to formulate, validate and/or discard several hypotheses at a time.

The system contains a Natural Language Interrogation subsystem that allows a biologist to formulate the temporal logic queries in (restricted) English.

GOALIE

GOALIE is a software system that uses the GO ontology biological process taxonomy (or any other ontology or controlled vocabulary collection – e.g. MeSH, UMLS, etc.) to automatically extract temporal invariants from numerical data organized along time (or concentration, dosage or other independent variable or combination thereof). The key contribution afforded by GOALIE is integrating data-driven reasoning about time course data-sets with model-building capabilities through the concept of *redescription* [25].

A *redescription* is a shift-of-vocabulary, or a different way of communicating a given aspect of information. Redescription mining is a technique to find sets (here, of genes) that afford multiple definitions. The inputs to redescription mining are the universal set of open reading frames (ORFs) in a given organism, and various subsets (called *descriptors*) defined over this universal set. These subsets could be based either on prior biological knowledge or defined by the outputs of algorithms operating on gene expression data. Example descriptors can be: ‘genes localized in cellular compartment nucleus’, and ‘genes involved in glucose biosynthesis.’ The goal of redescription mining is to connect these diverse vocabularies, by relating set-theoretic constructs formed over the descriptors.

In the first step of the implemented method, GOALIE analyzes a time-course micro-array experiment by weighing time-points with a sliding-window (in the simplest case using a Haar square weight function, or, in a more interesting case a smoother function, e.g. a Gaussian). The approach has its roots in previous work done by the NYU Courant Bioinformatics Group on multi-frequency analysis of signals. The result of this first step (which can be carried out in several ways) is a set of (overlapping) windows.

In a second step, GOALIE leverages the growing body of controlled vocabulary (ontological) annotations for genes and proteins by constructing *redescriptions* of each cluster in each window. Each cluster in each “window” will eventually get associated with a number of “terms” from the controlled vocabulary (e.g. from the GO process taxonomy). This association is done by performing different data-dependent statistical tests: simple implication covering based on Jaccard similarity, Hypergeometric test, Fischer Exact Test, along with appropriate statistical corrections like Bonferroni and Benjamini-Hochberg, to reduce the false discovery rate of inferred re-descriptions. This construction for a fixed set of clusters has also been explored previously, but in GOALIE this idea is coupled with the time-course analysis of numerical measurements, thus bringing to bear the correlation among processes happening within a biological system. The approach just described is completely novel.

In the third step, GOALIE creates a set of graph relationships between windows based on the associations among clusters and terms from the controlled vocabulary. This set of graph-relationships is the basis for the construction of temporal logic formulae describing the biological system at a phenomenological level (see below: the fifth step). The construction of this graph is straightforward but it strongly depends on the choice of controlled vocabulary or ontology, on the quality of the basic annotations available (e.g., annotation of a given gene product with a number of terms), and on the quality of the statistical tests used in the previous step.

Finally, in the fourth step, the set of graph relationships is organized in a directed acyclic graph (although circularities can be re-introduced by a wrapping technique). An edge is placed between a cluster in a window and another cluster in a previous or successor window. Each edge is tagged with the terms that are shared between the two clusters re-descriptions. Each edge is also tagged with the terms that are associated only to the first cluster, and with the terms that are associated only to the second cluster. The set of temporal logic sentences is reconstructed by analyzing different "chains" of edges in the DAG. For example, finding a set of terms that appear in each edge of a chain from the initial window to the last window generates a particular temporal logic sentence, denoting the invariance of the set of terms.

GOALIE has been described in [26, 27]

NYUMAD/NYUSIM

NYUMAD is a database for storing micro-array data based on the MAML model definition [28]. NYUSIM is a database system for storing simulation data. The core system underneath both NYUMAD and NYUSIM is organized in a three-tier architecture ensuring scalability. A Postgreql relational database management system forms the back-end tier. The middle application tier comprises Java servlets and supporting modules that respond to client requests and interact with the database. The front-end is a Java application that provides an easy and intuitive GUI (graphical user interface). The GUI communicates with the server side using an XML data exchange format over HTTP. The architecture is illustrated in Figure 6.

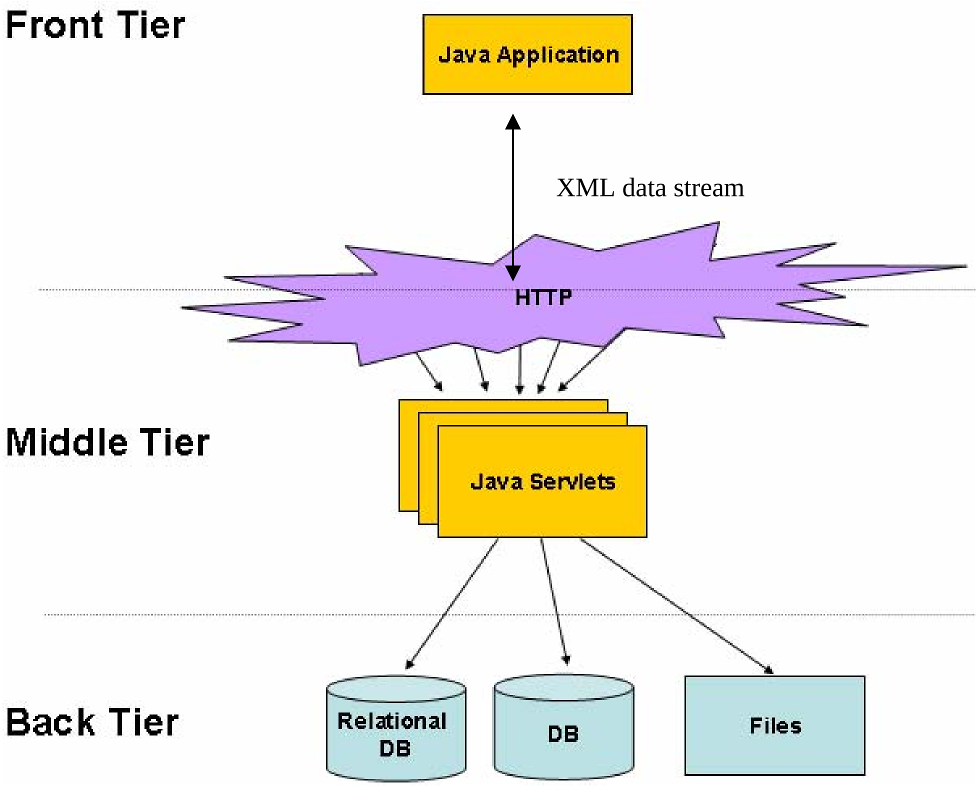
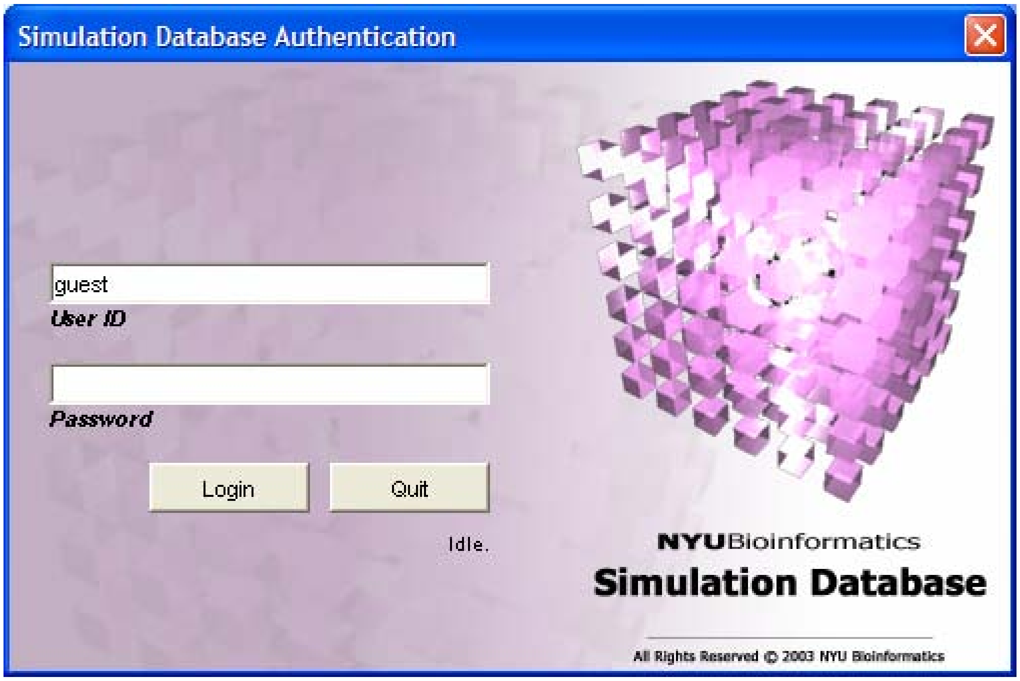


Figure 6 NYUMAD/NYUSIM three tier architecture.

The system is accessible to anyone with an Internet connection. See http://bioinformatics.cat.nyu.edu/nyumad for information on how to download and use the client GUI application. Users with IDs and passwords can save, edit and retrieve private data. Other users can log on as a ‘guest’ and view and retrieve public data. The ‘login’ screen is shown in the Figure 7.



Simulation Database Authentication

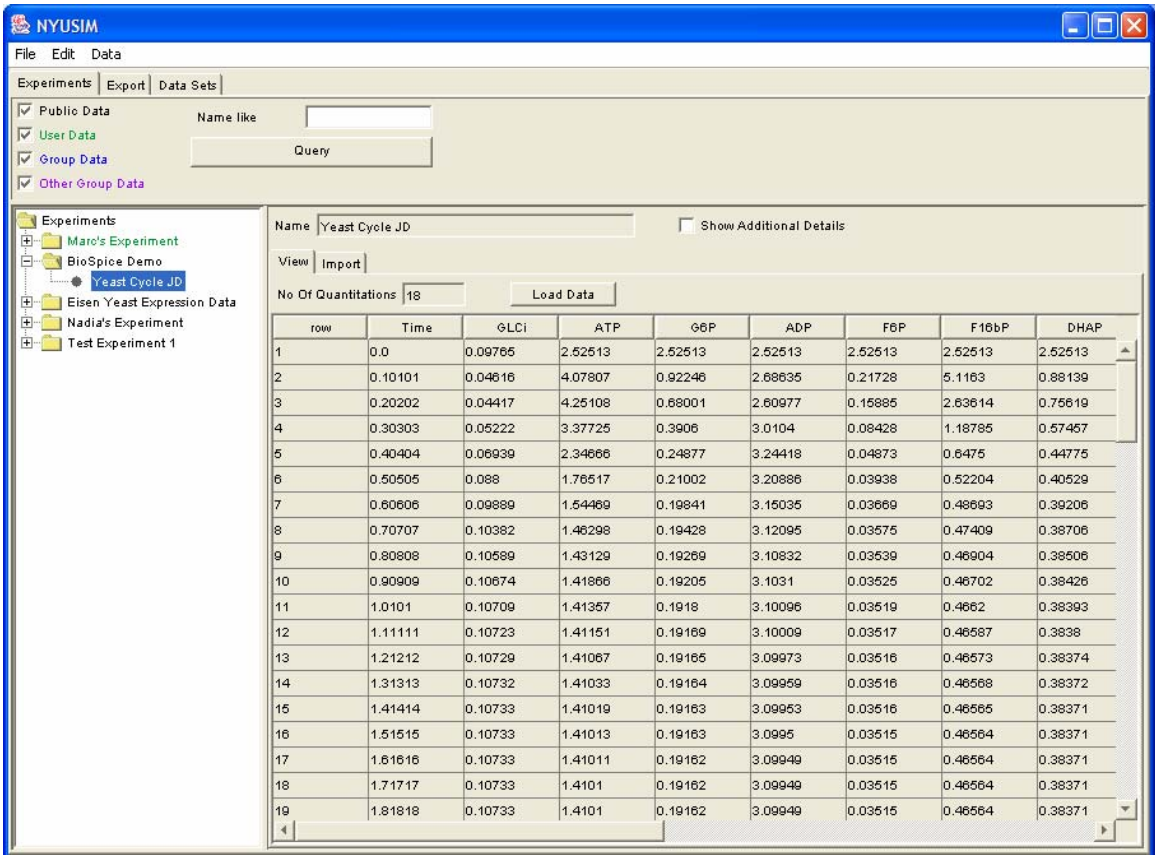
guest
User ID

Password

Figure 7 The NYUSIM client login window.

The system allows controlled access to data so that only users with the correct authorization can view private data. Each dataset has an ownership that determines its visibility. Collaborating groups can allow shared visibility of the data between their groups. After publication, data can be made publicly available with a simple command. Public data can be viewed by all users, including guest users.

The system stores a set of simulation trace data as a matrix, each column representing a simulated variable and each row representing a time point. Simulation data sets (matrices) are grouped under an experiment. Users create experiments, and for each experiment they can generate and store several sets of simulation data. Figure 8 shows a view of one such data set.



☑ Public Data
☑ User Data
☑ Group Data
☑ Other Group Data

Name like

Name Yeast Cycle JD

□ Show Additional Details

No Of Quantitations 18

row	Time	GLCi	ATP	G6P	ADP	F6P	F16bP	DHAP
1	0.0	0.09765	2.52513	2.52513	2.52513	2.52513	2.52513	2.52513
2	0.10101	0.04616	4.07807	0.92246	2.68635	0.21728	5.1163	0.88139
3	0.20202	0.04417	4.25108	0.68001	2.60977	0.15885	2.63614	0.75619
4	0.30303	0.05222	3.37725	0.3906	3.0104	0.08428	1.18785	0.57457
5	0.40404	0.06939	2.34666	0.24877	3.24418	0.04873	0.6475	0.44775
6	0.50505	0.088	1.76517	0.21002	3.20886	0.03938	0.52204	0.40529
7	0.60606	0.09889	1.54469	0.19841	3.15035	0.03669	0.48693	0.39206
8	0.70707	0.10382	1.46298	0.19428	3.12095	0.03575	0.47409	0.38706
9	0.80808	0.10589	1.43129	0.19269	3.10832	0.03539	0.46904	0.38506
10	0.90909	0.10674	1.41866	0.19205	3.1031	0.03525	0.46702	0.38426
11	1.0101	0.10709	1.41357	0.1918	3.10096	0.03519	0.4662	0.38393
12	1.11111	0.10723	1.41151	0.19169	3.10009	0.03517	0.46587	0.3838
13	1.21212	0.10729	1.41067	0.19165	3.09973	0.03516	0.46573	0.38374
14	1.31313	0.10732	1.41033	0.19164	3.09959	0.03516	0.46568	0.38372
15	1.41414	0.10733	1.41019	0.19163	3.09953	0.03516	0.46565	0.38371
16	1.51515	0.10733	1.41013	0.19163	3.0995	0.03515	0.46564	0.38371
17	1.61616	0.10733	1.41011	0.19162	3.09949	0.03515	0.46564	0.38371
18	1.71717	0.10733	1.4101	0.19162	3.09949	0.03515	0.46564	0.38371
19	1.81818	0.10733	1.4101	0.19162	3.09949	0.03515	0.46564	0.38371

Figure 8 A screenshot of the NYUSIM interface. The main window shows a time course dataset obtained by running the Simpathica simulation front end.

The GUI makes the importing of new data easy. New data sets are imported into the system by cutting and pasting into an importing area or by loading from a file. After importing data, synthetic data sets can be created by combining columns from different but compatible matrices. Data can be exported to the system clipboard from all the screens where matrix data is loaded or viewed, providing very flexible and efficient data retrieval for further analysis. There is a custom 'Export' screen where any combination of compatible columns can be exported.

The security model of the system controls visibility and read/write access to the data. Each user belongs to a primary group that gives them read access to all the data belonging to members of that group. An administrator tool is used to set and edit a user's write access and additional access rights to data from other groups.

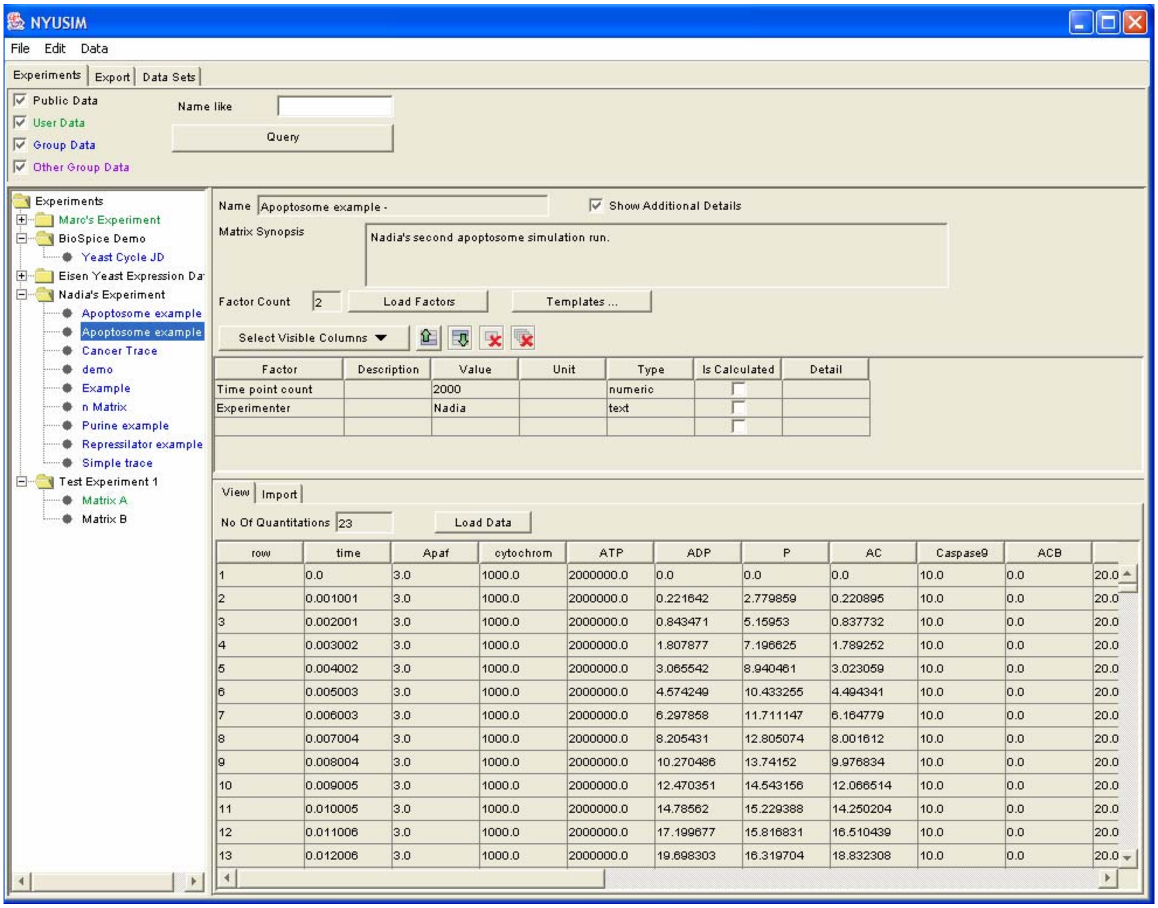
For viewing data, users have the flexibility to restrict data query to data categories of interest. This will be a useful feature as the number of experiments and data sets increases. Different query and response panels can be seen in Figure 8 and Figure 9. There are four major data categories:

1. *Public Data* – visible to all users including 'guest' users
2. *User Data* – the user's private data, visible only to other members of the same group

3. *Group Data* – data from other members in the same group as the user
4. *Other Group Data* – data from other groups giving the user access rights

Collaborating groups sharing data will see the data from other groups under the 'Other Group Data' category. In the tree view of the data hierarchy, the different data categories are color-coded for easy identification. In addition, the data query can be restricted to experiments with names matching a given pattern.

In addition to basic simulation data, it is possible to store associated data such as experimental factors and parameters as well as free format descriptive text for each experiment or data set. If there are common sets of factor and parameter data, a template of such factors can be created for easy input. Figure 9 shows a data set with two factors and a very brief synopsis.



☑ Public Data
☑ User Data
☑ Group Data
☑ Other Group Data

Name like

Name Apoptosome example -

☑ Show Additional Details

Matrix Synopsis Nadia's second apoptosome simulation run.

Factor Count 2

Factor	Description	Value	Unit	Type	Is Calculated	Detail
Time point count		2000		numeric	□	
Experimenter		Nadia		text	□	
					□	

No Of Quantitations 23

row	time	Apaf	cytochrom	ATP	ADP	P	AC	Caspase9	ACB	
1	0.0	3.0	1000.0	2000000.0	0.0	0.0	0.0	10.0	0.0	20.0
2	0.001001	3.0	1000.0	2000000.0	0.221642	2.779859	0.220895	10.0	0.0	20.0
3	0.002001	3.0	1000.0	2000000.0	0.843471	5.15953	0.837732	10.0	0.0	20.0
4	0.003002	3.0	1000.0	2000000.0	1.807877	7.196625	1.789252	10.0	0.0	20.0
5	0.004002	3.0	1000.0	2000000.0	3.065542	8.940461	3.023059	10.0	0.0	20.0
6	0.005003	3.0	1000.0	2000000.0	4.574249	10.433255	4.494341	10.0	0.0	20.0
7	0.006003	3.0	1000.0	2000000.0	6.297858	11.711147	6.164779	10.0	0.0	20.0
8	0.007004	3.0	1000.0	2000000.0	8.205431	12.805074	8.001612	10.0	0.0	20.0
9	0.008004	3.0	1000.0	2000000.0	10.270486	13.74152	9.976834	10.0	0.0	20.0
10	0.009005	3.0	1000.0	2000000.0	12.470351	14.543156	12.066514	10.0	0.0	20.0
11	0.010005	3.0	1000.0	2000000.0	14.78562	15.229388	14.250204	10.0	0.0	20.0
12	0.011006	3.0	1000.0	2000000.0	17.199677	15.816831	16.510439	10.0	0.0	20.0
13	0.012006	3.0	1000.0	2000000.0	19.698303	16.319704	18.832308	10.0	0.0	20.0

Figure 9 NYUMAD/NYUSIM interface used to add experimental factors and other information to a given dataset.

4. Conclusions

In conclusion, we note that while the application of the systems approach to biology is relatively new, we have made significant strides rather quickly by crosscutting with many important tools and techniques from control theory and computer science: discrete and hybrid automata, non-linear system analysis, Kripke models, temporal logic model checking, modeling databases, languages and environments, etc. The NYU Bioinformatics Group, in collaboration with all other research groups in the DARPA BioCOMP program, has made many significant contributions in each of these areas through techniques, theories and tools. We believe these results will form the foundations for the new science of Systems Biology.

5. References

[1] R. Alur, C. Belta, F. Ivančić, V. Kumar, M. Mintz, G. H. Pappas, H. Rubin, and J. Shug, "Hybrid Modeling of Biomolecular Networks," in *Proceedings of the 4th International Workshop on Hybrid Systems: Computation and Control*, vol. 2034, *LNCS*: Springer, 2001.

[2] M. Antoniotti, I. T. Lau, and B. Mishra, "Naturally Speaking: A System Biology Tool with a Natural Language Based Interfaces," in *Biological Language Conference*. Carnegie Mellon University, Pittsbugh, PA, U.S.A., 2004.

[3] M. Antoniotti, B. Mishra, C. Piazza, A. Policriti, and M. Simeoni, "Taming the Complexity of Biochemical Models through Bisimulation and Collapsing: THeory and Practice," *Theoretical Computer Science*, vol. 325, pp. 45-67, 2004.

[4] M. Antoniotti, F. Park, A. Policriti, and N. Ugel, "Foundations of a Query and Simulation System for the Modeling of Biochemical Processes," in *Pacific Symposium on Biocomputing (PSB03)*, R. B. Altman, A. K. Dunkler, L. Hunter, T. A. Jung, and T. E. Klein, Eds.: World Scientific, 2003, pp. 116-127.

[5] M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra, "xS-systems: eXtended S-systems and Algebraic Differential Automata for Modeling Cellular Behaviour," in *High Performace Computing (HiPC 2002)*, vol. 2552, *LNCS*: Springer Verlag, 2002, pp. 431-442.

[6] M. Antoniotti, A. Policriti, N. Ugel, and B. Mishra, "Model Building and Model Checking for Biochemical Processes," *Cell Biochemistry and Biophysics*, vol. 38, pp. 271-286, 2003.

[7] A. Casagrande, C. Piazza, and B. Mishra, "Semi-Algebraic Constant Reset Hybrid Automata - SACoRe," in *44th IEEE Conference on Decision and Control (CDC05)*. Seville, Spain, 2005.

[8] B. Mishra, "A Symbolic Approach to Modelling Cellular Behaviour," in *High Performance Computing (HiPC 2002)*, vol. 2552, *LNCS*, S. Sanhi, V. K. Prasanna, and U. Shukla, Eds.: Springer Verlag, 2002, pp. 725-732.

[9] B. Mishra, M. Antoniotti, S. Paxia, and N. Ugel, "Simpathica: A Computational Systems Biology Tool within the Valis Bioinformatics Environment," in *Computational Systems Biology*, E. Eiles and A. Kriete, Eds.: Elsevier, 2005.

[10] B. Mishra and A. Policriti, "Systems Biology and Automata," in *3rd Workshop on Computation of Biochemical Pathways and Genetic Networks*. Villa Bosch, Heidelberg, Germany: Springer Verlag, 2004.

[11] B. Mishra and A. Policriti, "Systems Biology, Automata, and Languages," in *Bioinformatics Italian Society Meeting (BITS 2004)*. Padua, Italy, 2004.

[12] V. Mysore and B. Mishra, "Algorithmic Algebraic Model Checking III: Approximate Models," in *INFINITY 05, Satellite Workshop of CONCUR 05*. San Francisco, CA, U.S.A., 2005.

[13] V. Mysore, C. Piazza, and B. Mishra, "Algorithmic Algebraic Model Checking II: Decidability of Semi-Algebraic Model Checking and its Applications to Systems Biology," in *Automated Technology for Verification and Analysis*. Taipei, Taiwan, 2005.

[14] C. Piazza, M. Antoniotti, V. Mysore, A. Policriti, F. Winkler, and B. Mishra, "Algorithmic Algebraic Model Checking I: Challenges from Systems Biology," in *17th International Conference on Computer Aided Verification (CAV05)*. The University of Edinburgh, Edinburgh, Scotland, U.K., 2005.

[15] C. Piazza and B. Mishra, "Stability of Hybrid Systems and Related Questions from Systems Biology," in *Advances in Control, Communication Networks and Transportation Systems: In Honor of Pravin Varaiya*, E. H. Abed, Ed.: Birkhauser, Boston, MA, U.S.A., 2005.

[16] L. S. do Carmo, C. Cummings, V. R. Linardi, R. Souza Dias, J. M. de Souza, D. A. dos Santos, J. W. Shupp, R. C. Peres Pereira, and M. Jett, "A Case Study of a Massive Staphylococcal Food Poisoning Incident," *Foodborne Pathogens and Disease*, vol. 1, pp. 241-246, 2004.

[17] R. Hammamieh, S. Bi, S. Mani, N. Chakraborty, C. Mendis, R. Das, and M. Jett, "Genetic variations in peripheral blood mononuclear cells in piglets used as an animal model for staphylococcal enterotoxin exposures," *Omics*, vol. 7, pp. 401-409, 2003.

[18] R. Hammamieh, S. Bi, R. Neill, N. Chakraborty, C. Mendis, R. Das, and M. Jett, "Global responses in gene expression responses in peripheral blood mononuclear cells to SEB: in vitro vs. in vivo modeling," *Biosensors and Bioelectronics*, vol. 20, pp. 719-727, 2004.

[19] Y. von Gessell, S. Mani, S. Bi, R. Hammamieh, J. W. Shupp, R. Das, G. D. Coleman, and M. Jett, "Functional Piglet Model for the Clinical Syndrome and Post Mortem Findings Correlated with Gene Expressions Responses Induced by Staphylococcal Enterotoxin B," *Experimental Biology and Medicine*, vol. 229, pp. 1061-1071, 2004.

[20] P. E. Barbano, M. Spivak, J. Feng, M. Antoniotti, and B. Mishra, "A Coherent Framework for Multi-resolution Analysis of Biological Networks with Memory: RAS Pathway, Cell Cycle and Immune System," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 102, pp. 6245-6250, 2005.

[21] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335-338, 2000.

[22] C. C. Guet, M. B. Elowitz, W. H. Hsing, and S. Leibler, "Combinatorial synthesis of genetic networks," *Science*, vol. 296, pp. 1466-1470, 2002.

[23] M. Antoniotti, P. E. Barbano, W. Casey, J.-W. Feng, N. Ugel, and B. Mishra, "Mutliple Biological Model Classification: From Systems Biology to Synthetic Biology," *BioCONCUR 04, Transactions on Computation Systems Biology*, 2005.

[24] E. O. Voit, *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists*: Cambridge University Press, 2000.

[25] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. Helm, "Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions," in *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, WA, U.S.A., 2004.

[26] M. Antoniotti, N. Ramakrishnan, and B. Mishra, "Reconstructing Formal Temporal Models of Cellular Events using the GO Process Ontology," in *Proceedings of the Eight Annual Bio-Ontologies Meeting, Satellite Workshop of ISMB05*. Detroit, MI, U.S.A., 2005.

[27] M. Antoniotti, N. Ramakrishnan, and B. Mishra, "GOALIE, A Common Lisp Application to Discover Kripke Models: Redescribing Biological Processes from Time Course Data," in *International Lisp Conference*. Stanford, CA, U.S.A., 2005.

[28] M. Rejali, M. Antoniotti, V. Cherepinsky, C. Leventhal, S. Paxia, A. Rudra, J. West, and B. Mishra, "Design and Implementation of a Versatile MicroArray Data Base," presented at CGC, Baltimore, MD, U.S.A., 2001.